Use of a novel evolutionary algorithm for genomic selection

Julie Hamon, Gaël Even, Romain Dassonneville, Julien Jacques, Clarisse Dhaenens

RESEARCH

Use of a novel evolutionary algorithm for genomic selection

Julie Hamon[1,2*], Gaël Even[2], Romain Dassonneville[2], Julien Jacques[1,3] and Clarisse Dhaenens[1,4]

[*]Correspondence:
julie.hamon1@gmail.com
[1]Inria Lille - Nord Europe, Lille, France
[2]Gènes Diffusion, 3595 rte de Tournai, Douai, France
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: In the context of genomic selection in animal breeding, an important objective is to look for explicative markers for a phenotype under study. The challenge of this study was to propose a model, based on a small number of markers, to predict a quantitative trait. To deal with a high number of markers, we propose using combinatorial optimization to perform variable selection, associated with a multiple regression model in a first approach and a mixed model in a second, to predict the phenotype.

Results: The efficiency of our two approaches, the first assuming that animals are independent and the second integrating familial relationships, was evaluated on real datasets. This reveals the importance of taking familial relationships into account as the performances of the second approach were better. For example, on PIC data the correlation is around 0.15 higher using our approach taking familial relationships into account than with the Lasso bounded to 96 selected markers. We also studied the importance of familial relationships on phenotypes with different heritabilities. Finally, we compared our approaches with classic approaches and obtained comparable results, sometimes better.

Conclusion: This study shows the relevance of combining combinatorial optimization with a regression model to propose a predictive model based on a reasonable number of markers. Although this implies more parameters to be estimated and, therefore, takes longer to execute, it seems interesting to use a mixed model in order to take familial relationships between animals into account.

Keywords: genomic selection; combinatorial optimization; regression

1 INTRODUCTION

Genomic selection of animal breeding deals with a genetic evaluation of animals from their DNA (extracted using biological samples such as blood or hair, or biopsy), based on markers covering the whole genome. Important insight in this domain is gained by establishing predictive models using genomic information. High-throughput genotyping data are analyzed in this study and an important feature of these data is the huge number of markers (p) compared to the number of subjects (n). So, in order to predict a quantitative trait using these data, the classic statistical problem of high dimensional regression ($n < p$) has to be solved.

Various methods have been proposed, including approaches based on best linear unbiased prediction (BLUP), Bayesian approaches or shrinkage regression methods. The choice of which method to use usually depends on the genetic architecture of the trait studied [1]. Indeed, for a given trait, if the distribution of effects is known to be normal, it is preferable to use a method such as G-BLUP while if the trait depends on areas of the genome with large effects, Bayesian methods are preferred. The challenge of this study was to find a predictive model based on a small number of markers allowing the selection of the best animals for a given phenotype, in order to produce small size chips for the phenotype under study. Indeed, low density chips are cheaper and it can be interesting, for example, to genotype a large amount of animals with this type of chip and genotype only the best one with a high density chip.

The problem of variable selection among a huge amount of variables can be seen as a combinatorial problem [2]. We therefore proposed dealing with this problem by using a combinatorial optimization approach. Modeling this problem as a combinatorial optimization problem is interesting as it allows efficient methods which have been developed for this kind of problem to be adopted. Here, the size of the problem is very large (it depends on the number of markers), hence a complete enumeration will not be possible. In this context, heuristic optimization approaches will be used. Such methods have been applied for variable selection in various domains, especially on microarray data or SNPs data in classification contexts. However, they can be adapted to a regression problem to deal with quantitative traits such as milk production or meat quality.

Among combinatorial optimization methods, metaheuristics are approximate algo-

rithms that can efficiently explore a very large search space in order to obtain a satisfactory solution [3]. In this study we adopted evolutionary algorithms, which are population based metaheuristics, based on Darwin's theory of evolution [4].
For this study, we suggested adressing the problem of variable selection in a high dimensional regression context by combining a combinatorial optimization approach for selecting subsets of variables and a statistical model to evaluate this subset. An interesting outcome is that the proposed algorithm affords the possibility of including familial relationships. Hence, to carry out experiments, real datasets from beef cattle and pigs were used to compare the proposed method with classic approaches for traits with various heritabilities.

2 MATERIEL AND METHODS

2.1 Data

In this study, cattle and pig data are used. Cattle data come from the Qualvigène project [5] in which Gènes Diffusion (`www.genesdiffusion.com`) is involved.

This program includes Charolais bulls and young bulls with 48 sires and $1,114$ bulls. The trait studied was the carcass yields with high heritability ($h^2 = 0.54$). Following pre-treatment on available animal data (including the removal of non-phenotyped animals for that trait), we finally obtained $1,107$ animals (48 sires) genotyped in 54K. Following quality control of the genotyping data, $43,896$ SNPs were retained for the study. We obtained an SNP data matrix size of $1,107 \times 43,896$, associated with a vector of size equal to $1,107$ for carcass yield. Values of the trait studied here were corrected for environmental effects and form the deregressed proofs [6]. To complete this data, pedigree information on $4,741$ animals was known.
The second dataset used is a pig dataset that PIC (a Genus company) has made available [7]. The dataset consisted of $3,534$ animals genotyped on the PorcineSNP60 chip (64,233 markers). These genotypes were filtered for a Minor Allele Frequency (MAF) > 0.001 and a proportion of missing genotypes by SNPs $> 10\%$. Markers on the X and Y chromosomes were excluded, yielding 52,842 SNPs. Pedigree information was also available, including parents and grandparents of the genotyped animals ($n = 6,473$). Genotyped animals had phenotypes for five traits, with heritability ranging from 0.07 to 0.62. The authors state that, "Each phenotype was

either corrected for environmental factors (e.g. year of birth or farm) and rescaled by correcting for the overall mean (traits 3, 4 and 5) or was a rescaled, weighted mean of corrected progeny phenotypes (traits 1 and 2), for which many animals have no individual performance data" [7].

2.2 Model

The objective was to predict a quantitative trait from a subset of quantitative variables. This can be modeled as a multiple linear regression, which we propose formulating as follows:

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + e_i, \quad i = 1, \ldots, n, \\ \mathbf{y} &= X(\boldsymbol{\beta} \cdot \boldsymbol{\gamma}) + \boldsymbol{e}, \end{aligned} \tag{1}$$

where

- $\mathbf{y}$ is a vector of dimension n (number of animals) representing the quantitative trait of interest,
- X are the fixed effects with $X = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ a $n \times (p+1)$ matrix with p the number of SNPs studied, the first column of X contains a 1
- $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_p)$ the coefficients to be estimated.
- $\gamma_j = (\gamma_1, \ldots, \gamma_p)^t$ equals 1 if the SNP j is in the model, 0 otherwise. The operator $\cdot$ corresponds to the product of the term-by-term vectors.
- $\boldsymbol{e}$ are Gaussian residuals assumed to be independent and identically distributed (i.i.d.) with zero mean and variance σ_e^2.

Parameters $\boldsymbol{\gamma}$, $\sigma^2_{\mathbf{e}}$, β_0 and $\{\beta_j : \gamma_j = 1, 1 \leq j \leq p\}$ have to be estimated.

In this proposed modeling, as in many approaches from the literature, animals are considered to be independent.

However, unlike human studies, familial relationships exist between individuals; these are described by means of a deep pedigree. This is important information, which must be taken into account to avoid, for example, considering SNPs as significant when they are not, and thereby increasing the number of false positives. We proposed integrating these familial relationships, using the pedigree, through a linear mixed model based on equation (1). A term $Z\mathbf{u}$ is added to this equation in order to introduce correlations between observations. This leads to equation (2).

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \sum_{k=1}^{q} z_{ik} u_k + e_i, \quad i = 1, \ldots, n,$$
$$\mathbf{y} = X(\boldsymbol{\beta} \cdot \boldsymbol{\gamma}) + Z\mathbf{u} + \boldsymbol{e}, \tag{2}$$

where $Z\mathbf{u}$ are the random effects representing familial relationships (animal model). These effects serve to reduce the number of false positive SNPs detected due to familial relationships [8].

The objective here was to estimate the parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_p)$ and $\gamma = (\gamma_1, \ldots, \gamma_p)$. As γ is a discrete parameter belonging to $\{0, 1\}^p$, determining the γ_j values is equivalent to determining variables that participate in the regression model. This problem is a typical feature selection problem, well known in data-mining, and which may be seen as a combinatorial problem. Hence it can be addressed using combinatorial optimization methods. In what follows, such a method is proposed for this task.

2.3 Validation

We compared the proposed approach with two classic regression methods used in genomics considering SNPs as fixed effects (as in our approach): *elastic net* [9] and *Lasso* [10]. These methods are shrinkage regression approaches, meaning that they shrink regression coefficients toward 0, which leads to select variables. The main difference with our approach is that they cannot take familial relationships into account.

Since the objective was to create a low density chip, we fixed the maximum number of markers selected by using our approach to 96, i.e. a classic low density chip size. We also compared our approach with the two previous methods bounded to 96 selected markers.

On the Qualvigene dataset, 100 young bulls were selected to constitute a validation set, leading to a training sample made up of 1,007 individuals. In order to generalize the results obtained, this split was performed 30 times, each time differently (generating 30 instances). We evaluated the performance of the two proposed

models: the first with a multiple linear regression (eq. (1)) and the second with a mixed model integrating familial relationships (eq. (2)).
We ran each method on the 30 generated instances. Results are evaluated both in terms of RMSEP (root mean square error of prediction - to minimize) on the validation set and in terms of correlation (to maximize) between the estimated trait and the real trait.

On the pig dataset, 100 subjects were selected (from the 3,534) to form the validation set and this split was performed differently 10 times in order to obtain 10 different instances. We studied the performance of the approaches for the five traits. As the results for both methods, elastic net and Lasso were similar, we present only those obtained using Lasso.

Our approach was developed in C++ using the PARADISEO platform [11] (`http://paradiseo.gforge.inria.fr/`), for metaheuristics setting. For classic approaches (elastic net, Lasso), we used R software with the *"glmnet"* procedure. The λ parameter for both methods was determined using *"cv.glmnet"* and the α parameter for elastic net by a 3-fold cross-validation. In our approach, the evaluation of a variable selection with a mixed model was performed using the BLUPF90 program in FORTRAN by Mistzal [12].
The results presented were computed at the regional cluster financed by Lille 1 University, the CPER Nord-Pas-de-Calais/FEDER, France Grille and CNRS.

3 OPTIMIZATION APPROACH

Evolutionary algorithms are search methods based on natural evolution [13] and the most popular one, used in this study, is the genetic algorithm [14]. The objective in this study was to search for a relevant subset of variables (markers) among a large amount of possible subsets.
Figure 1 shows the general scheme of the algorithm.
It starts with the initialization of a population of n individuals where an individual is an encoding version of a candidate solution (in our study a solution describes a subset of variables). Each solution is evaluated and $n/2$ couples of solutions are selected. Each couple generates two new solutions through the crossover operator

and then a mutation operator might be applied in order to diversify these new solutions. A replacement strategy chooses, among initial solutions and new solutions, the solutions of the next population. These successive steps correspond to a generation. The algorithm stops when it reaches a given stopping criterion. Different selection, crossover, mutation and replacement operators or stopping criteria may be used. We present the choices we made below.

3.1 Encoding of a solution

The representation of a solution plays a major role in the implementation of a metaheuristic since it influences the choice of the operators and the evaluation function. We chose to use a binary vector indicating whether a variable is selected (1) or not (0) since it is very close to the statistical models (1) and (2) presented above (this is equivalent to the vector $(\gamma_1, \ldots, \gamma_p)$). In addition, this encoding provides a simple but effective design of the neighborhood. Example of a solution:

1	0	0	1	1	0	1	0

In this solution, variables 1, 4, 5 and 7 are selected. The size of a solution (8 in this case) is equal to the total number p of variables in the dataset studied.

3.2 Objective function

The aim of the optimization method is to effectively explore a large search space matching all possible subsets of variables. Therefore, this method uses an evaluation criterion (fitness function) able to associate one quality measure with each solution. In our context, the objective was to identify the best subset of variables, in other words, the one which will provide the best predictive model. A well-known difficulty in data-mining is how to assess the model's ability to predict a trait from data that were not used to develop the model (validation sample). The objective function used, depending on the model considered (a multiple linear regression model or a mixed model) will be described later.

3.3 Initialization

The classic initialization of solutions of an evolutionary algorithm is to set solutions randomly (in a uniform manner). As the representation of a solution is a binary vector, the purpose here is to set up each bit to 0 or 1. To obtain diversified initial solutions, we wish to have solutions of different sizes (with different numbers

of selected variables), while remaining below the maximum allowable number of variables (96 variables here). Therefore, for each solution, its number k of selected variables is uniformly chosen in a predefined interval $[min., max.]$. Moreover, to accelerate convergence, and in order to obtain interesting initial solutions, we compared pure random initialization with guided initializations. Three configurations have been tested.

- The first consists in uniformly choosing the k variables of each solution of the initial population.
- The second configuration consists in initializing all the solutions of the initial population using the variables selected by the Lasso method. Indeed, the Lasso method (not limited in number of selected variables) allows us to obtain a subset of *a priori* interesting variables. So, for each solution (individual) of the initial population, we uniformly selected variables among those obtained by the Lasso method. If the number k of variables desired for the solution is greater than the number of variables extracted by the Lasso method, we choose all variables identified by the Lasso method, and add variables uniformly selected from among the others.
- The third configuration consists in combining the two described above. To do this, we separate the initial population into two parts. For the first half, solutions of the initial population are randomly generated, while solutions for the second half are constructed using variables selected by Lasso.

Experimentations on simulated data [15] show that initializations based on the asso method (configurations 2 and 3 presented above) give better results than a pure uniform initialization. As there was no significant difference between the two configurations based on Lasso method, we chose to use the third configuration, that is an initialization based on Lasso for 50% of the solutions of the initial population, in order to maintain diversity.

3.4 Selection

The selection process of an evolutionary algorithm aims to determine the individuals that will breed and how many children each couple will generate. This is equivalent to determining the subset of variables which will be used for the creation of new

subsets. Several selection strategies are possible including roulette wheel selection, stochastic universal sampling or tournament selection [3]. We chose tournament selection as it does not converge too fast and also helps to maintain diversity. Tournament selection consists in randomly selecting m individuals, m being the size of the group tournament. The best individual among the m individuals will be retained. The selection of n individuals requires n executions of a tournament.

3.5 Reproduction

Once the parents are selected, the reproduction phase applies variation operators such as crossover and mutation to generate children. The choice of binary encoding of solutions would allow us to use classic crossover and mutation operators [16]. Nevertheless, the choice of **crossover** operator may depend on the problem studied, in order to ensure an efficient one. Indeed, in the context of feature selection, traditional operators such as 1-point or 2-point crossovers may have a negative effect since they may "break" some interesting blocks. Therefore, we chose to use a crossover operator adapted to the problem of feature selection, the Subset Size-Oriented Common Feature (SSOCF [17]). The principle is described in Figure 2. Variables in common to both parents are kept by the children. The others are inherited from the parents with the probability $(n_i - n_c)/n_u$ where n_i is the number of variables selected by the i^{th} parent, n_c is the number of variables selected jointly by both parents and n_u the number of variables unshared by the parents (variables selected by one of the parents, but not both). The objective of this method is, on the one hand, to keep the blocks of useful information and on the other hand, to keep for the children the variables shared by their parents.

The **mutation** is a unary operator (one input solution) applied to an individual to change it slightly. In a binary representation of solutions, the mutation typically used is a bit-flip. Two types of mutation were used in our algorithm based on the number of selected variables in the current solution:

- flip a (small) percentage of bits uniformly determined among all variables when the number of variables in the selected current solution is less than the maximum desired number of variables ($\Rightarrow$ addition or deletion of variables).

- flip a (small) percentage of bits uniformly determined among selected variables (bit = 1) when the maximum number of desired variables is reached ($\Rightarrow$ deletion of variables).

At the reproduction step, crossover and mutation are not applied consistently. Indeed, the crossover rate is used to define the probability that two selected parents are crossed to generate children. Similarly, the mutation rate is the probability of applying a mutation to a solution. We compared the performances of the algorithm on simulated data [15] using low (0.2) and high (0.8) crossover and mutation rates. We finally chose to keep a crossover rate of 0.8 and a mutation rate of 0.8.

3.6 Replacement

The population size must be constant over generations. Hence, when children are generated, all parents and children cannot be kept. The replacement procedure, the last step of a generation, will help to define the survivors among parents and children generated. The replacement procedure that we chose here was to keep a child only if it is better than the worst of the remaining parents. When a child is preserved, the worst parent is deleted. The worst parents are replaced progressively by the best children.

3.7 Stopping criteria

The evolutionary algorithm is an iterative approach for which it is necessary to set a stopping criterion. Here, we set a maximal number of generations, determined empirically depending on the evolution curve of the best solution of the population.

3.8 Diversification

During the evolution of the evolutionary algorithm, a failure that can be observed is the stagnation of the search. To avoid this, diversification methods are proposed such as the stochastic diversity of migration or "Random Immigrant" [18]. The idea is to replace a portion of the population by individuals generated uniformly when the best individual of the population has not been improved for a given number of generations. In our algorithm, when the best individual of the population does

not change for a fixed number of generations, all individuals whose fitness is lower than the average fitness of the population are replaced by new individuals uniformly generated.

3.9 Parallelization

During the evolutionary algorithm, for a generation, several solutions (children generated) have to be evaluated. The evaluation may take time as a regression (computation of the coefficients of each marker) has to be performed. Thus, to reduce execution time, we proposed making these evaluations in parallel. We therefore implemented a synchronous parallel version of the algorithm with the SMP module of PARADISEO [11]. The aim is to parallelize, at every generation, the evaluations of children (solutions) of the evolutionary algorithm using the scheme "master / slave". Once all children are generated, their evaluations are independent so they are performed in parallel. During the evaluation phase, the master sends one solution to evaluate per slave and they send back the fitness of the solution received.

4 A STATISTICAL FITNESS FUNCTION

As we saw in Section 3.2, such an optimization method is based on a fitness function which evaluates the quality of solutions. The quality of a solution (a subset of variables) was assessed according to the quality of the underlying model (i.e. how best it fit the data). We defined a fitness function for each of the models proposed previously: multiple linear regression (1) and mixed model (2).

4.1 Multiple linear regression

Through multiple linear regression, a range of model selection methods is available in the literature (e.g. [19]). The most commonly used criteria are the AIC criterion (Akaike Information Criterion) [20], the BIC criterion (Bayesian Information Criterion) [21] and cross-validation. Unlike the AIC criterion, the BIC criterion tends to penalize complex models more heavily and therefore seems more appropriate to our objective of variable selection in high dimension. In a previous study, we compared three criteria [15]: BIC and two types of cross-validation (k-fold and leave-one-out) on simulated data and the BIC criterion gave the best results. We used it in this study.

4.2 Mixed model

For our second model, the quality of a solution was evaluated with a 3-fold cross-validation. Indeed, calculating the BIC requires calculating the likelihood of the model. However, the method of maximum likelihood is not suitable for mixed models and the use of restricted maximum likelihood (REML) is recommended for this type of model. An adaptation of the BIC has been proposed by [22] under repeated data but this is not the case of our data, so we chose to use 3-fold cross-validation.

5 RESULTS

In order to analyze performance of the proposed methods, we compared them on the two presented datasets using elastic net and Lasso approaches without and with a restriction on the number of selected markers. Figure 3 illustrates our results on the cattle data (Qualvigène project).

Our approach based on multiple linear regression (LM) allowed us to obtain results comparable to classic approaches bounded to 96 selected markers. Adding familial relationships using mixed models improved the results of our method from a correlation of 0.48 with LM to a correlation of 0.56 with MM. Moreover, this new approach outperformed classic approaches (the Student's test on the mean prediction error concluded with a significant difference between MM and EN96 or MM and L96) and became comparable to unlimited approaches (mean RMSEP equal to 0.49 for MM against 0.41 for the Lasso method (Las), for example).

Figures 4 to 8 illustrate results (RMSEP and correlation) for the five traits on the pig dataset. We observed that whatever the trait, our approach based on a multiple linear regression performed slightly better than Lasso limited to 96 selected markers. Moreover, performance is improved with our second approach including familial relationships so as to outperform the classic approach (the Student's test concluded with a significant difference between L96 and MM for all traits). For example, on the trait T1, the mean prediction error of the Lasso method limited to 96 variables selected was equal to 0.55; it decreased to 0.51 for our first approach and to 0.43 for our MM approach. On this trait, the prediction error of our last

approach outperformed that of the Lasso method (0.46).

In order to evaluate the influence of heritability on the performances of the different methods, results were compared on pig data on 5 traits with different heritabilities: 0.07, 0.16, 0.38, 0.58 and 0.62. For traits T2, T3 and T4, taking into account familial relationships improved the results (significant Student's test). However, although whatever the trait MM is always better than LM, sometimes the difference is small. For trait T1, the difference observed between LM and MM was significant in terms of RMSEP ($p - value = 0.03$) but not in terms of correlation ($p - value = 0.07$). For the trait T5, the difference between LM and MM was not significant.

The execution times of our approaches were slightly higher than those of classic approaches especially because they required the execution of the Lasso to be initialized. Table 1 shows the execution times for the different methods on the Qualvigène dataset.

We observed that the evaluation using a mixed model takes much longer than the multiple linear regression due to the high number of parameters to be estimated. Indeed, the actual execution time of the algorithm (once the initialization time was removed) with the mixed model was 7 minutes compared with 10 seconds for the linear regression. However, the execution times of our approaches were reasonable compared with the time taken to collect and pre-process data.

6 DISCUSSION

On the real datasets used, our approaches lead to similar or even better results than classic approaches. This enabled us to validate the relevance of combining a combinatorial optimization method and a regression to solve our problem.

We observed that methods unlimited in the number of selected markers (EN and Las) obtained the bests results (with a correlation of around 0.6). However, they selected too many variables ($\approx$ 580 and 300 respectively) and were not suitable for our problem. Indeed, selecting a large amount of variables results in a more accurate model, but our objective was to select a limited number of variables. The first model proposed in this study, based on a multiple linear regression (LM), assumes

that animals are independent. This is not the case in our data so we proposed the second model including familial relationships using a mixed model (MM). As the assumptions of LM are not met in our data but those of MM are, we expected to have better results with MM than with LM. This was confirmed by the results obtained on real cattle and pig data, which showed the importance of including familial relationships for these datasets. Regarding the results obtained on the pig dataset, as we have 5 traits with low to high heritability, we can measure the impact of the heritability of the trait on the performances of the different approaches. First, if we look at the results in terms of correlation (which are comparable from one trait to another), the methods performs better on low heritability traits. Moreover, if we compare LM and MM the difference in terms of correlation is not significant for the less heritable (T1) and the most heritable (T5) traits but significant for the others. It seems interesting, therefore, for this type of data, to integrate familial relationships for trait with moderate heritabilities but not necessarily for very low or high heritability.

In an evolutionary algorithm, it is difficult to fine-tune parameters. For each operator, we tested several possibilities (the most popular regarding this kind of data), evaluated their performance on simulated data and chose the best. Our approach is flexible regarding the statistical model used to evaluate subsets of variables. Indeed, we first performed a multiple linear regression and next a mixed model. However, this can be easily changed, for example by combining a multiple linear regression to start the search and a mixed model to refine the search. It could be also possible to test other approaches, such as Bayesian models.

In order to evaluate the quality of our approach, we extracted 100 bulls from our original datasets. We decided to choose these 100 bulls from among the young ones given our end objective of predicting performance on young bulls. Another outcome could be to extract a family if the objective was to predict the trait under study for an animal unrelated to those in the study.

Results were presented in terms of correlation and Root Mean Square Error of Prediction (RMSEP). Indeed, although the majority of genomic selection studies

present the results in terms of correlation, this measure is less accurate than the RMSEP. The RMSEP evaluates the difference between each prediction and real values whereas the correlation only looks at their distribution and evaluates whether they go the same way. So, if for all subjects the trait is estimated with a lower value than the real one, the correlation will be good whereas the RMSEP will be bad. If the objective is only to select the best animals, the correlation is a good indicator but if it is also interesting to have a good estimation for the trait, RMSEP is more accurate.

In genotyping data, some markers are in high Linkage Disequilibrium (LD). In our approach, if a marker is in high LD with another one already in the model, as it is not adding more information to the actual regression model, it is likelythat it will not be selected. In the final model, the LD between markers is low so that they explain different parts of the trait.

Our objective was to find a predictive model based on a small number of markers (96). This kind of very low density chip could be a decision tool for breeders in order to select animals to be genotyped on a 54K chip for example or in a 6K chip with imputation. Results on cattle and pig datasets showed that our approach obtains better results than elastic net or Lasso method in a reasonable computational time. Some may argue that a very low density chip (96 SNPs), specific to a given trait, became less interesting once imputation and "low" density (6K) chips were used. This may be true for cattle. But for livestock such as pigs or poultry, 96 SNP chips still appear to be interesting tools.

7 CONCLUSION

The objective of this study was to select a subset of relevant markers to predict a quantitative trait. We proposed a novel approach based on an evolutionary algorithm combined with a statistical model. We compared two statistical models, the first without familial relationships and the second integrating them using the pedigree information.

We first showed the importance of including of familial relationships in the statistical model as prediction on a validation set was better on the real datasets tested. Due

to its powerful exploration of the search space, the optimization approach makes it possible to find the SNPs of interest. Our approaches performed as effectively as the most efficient approaches used in the field and sometimes outperformed them.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JH performed the analysis and drafted the manuscript. JH, CD, JJ and GE designed the study and developed the method. GE and RD prepared phenotypic and genotypic data. CD, JJ, GE and RD revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We highly appreciate and thank the technical staff at the CRI-Lille 1 center for their strong and helpful support. We thank the ANR AGENAE and APIS-GENES for allowing us to use the Qualvigène data.

Author details

[1]Inria Lille - Nord Europe, Lille, France. [2]Gènes Diffusion, 3595 rte de Tournai, Douai, France. [3]Laboratoire Paul Painlevé / CNRS, Université Lille 1, Lille, France. [4]LIFL, Université Lille 1 , Lille, France.

References

1. Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E.: Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in holstein cattle as contrasting model traits. PLoS Genet **6**(9), 1001139 (2010)
2. Corne, D., Dhaenens, C., Jourdan, L.: Synergies between operations research and data mining: The emerging use of multi-objective approaches. European Journal of Operational Research **221**(3), 469–479 (2012)
3. Talbi, E.-G.: Metaheuristics. John Wiley & Sons, Inc., Hoboken, NJ, USA (2009)
4. Darwin, C.: On the Origin of the Species by Means of Natural Selection: Or, The Preservation of Favoured Races in the Struggle for Life. John Murray, London (1859)
5. Allais, S.: Détection et validation de marqueurs génétiques impliqués dans la qualité de la viande bovine. PhD thesis, AgroParisTech (January 2011)
6. VanRaden, P.M., Wiggans, G.R.: Derivation, calculation, and use of national animal model information. Journal of dairy science **74**(8), 2737–2746 (1991). PMID: 1918547
7. Cleveland, M.A., Hickey, J.M., Forni, S.: A common dataset for genomic analysis of livestock populations. G3: Genes|Genomes|Genetics **2**(4), 429–435 (2012)
8. Habier, D., Fernando, R.L., Dekkers, J.C.M.: The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**(4), 2389–2397 (2007)
9. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Statist. Soc. B **67 Part 2**, 301–320 (2005)
10. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1994)
11. Cahon, S., Melab, N., Talbi, E.-G.: ParadisEO: a framework for the reusable design of parallel and distributed metaheuristics. Journal of Heuristics **10**(3), 357–380 (2004)
12. Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D.H.: BLUPF90 and related programs (BGF90). In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production: August 2002, Montpellier, pp. 1–2 (2002)
13. Pal, S.K., Bandyopadhyay, S., Ray, S.S.: Evolutionary computation in bioinformatics: a review. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **36**(5), 601–615 (2006)
14. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence vol. viii. U Michigan Press, Oxford, England (1975)

15. Hamon, J.: Optimisation combinatoire pour la sélection de variables en régression en grande dimension : Application en génétique animale. PhD thesis, Université des Sciences et Technologie de Lille - Lille I (November 2013)
16. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, (1989)
17. Emmanouilidis, C., Hunter, A., MacIntyre, J.: A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. In: Proceedings of Congress on Evolutionary Computation, pp. 309–316 (2000)
18. Grefenstette, J.: Genetic algorithms for changing environments. In: Parallel Problem Solving from Nature 2: Amsterdam, pp. 137–144 (1992)
19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition, (2009)
20. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974)
21. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)
22. Delattre, M., Lavielle, M., Poursat, M.-A.: BIC selection procedures in mixed effects models (2012)

Figures

Figure 1 Evolutionary algorithm. General scheme of an evolutionary algorithm.

Figure 2 SSOCF. Subset Size-Oriented Common Feature : a crossover operator.

Figure 3 Performances on Qualvigène dataset. Boxplot evaluating performances of classical approaches (elastic-net (EN), lasso (Las)), classical approaches limited to 96 SNPs selected (EN96, L96) and our two approaches (based on multiple linear regression (LM) and on mixed model (MM)), in term of RMSEP (to minimize) on the left and of correlation (to maximize) on the right.

Figure 4 Performances on PIC, trait T1 ($h^2 = 0.07$).

Figure 5 Performances on PIC, trait T2 ($h^2 = 0.16$).

Figure 6 Performances on PIC, trait T3 ($h^2 = 0.38$).

Figure 7 Performances on PIC, trait T4 ($h^2 = 0.58$).

Tables

Figure 8 Performances on PIC, trait T5 ($h^2 = 0.62$).

Table 1 Execution time of different methods on cattle data

EN	Lasso	EN96	L96	LM	MM
35 min.	3 min.	16 min.	1 min.	3 min. 10	10 min.